論文 / 著書情報
Article / Book Information

論題(和文)	口唇の深度画像を用いたディープオートエンコーダによるマルチモーダル音声認識
Title(English)	
著者(和文)	安井勇樹, 岩野公司, 井上中順, 篠田浩一
Authors(English)	Yuki Yasui, Koji Iwano, Nakamasa Inoue, Koichi Shinoda
出典(和文)	情報処理学会研究報告 SLP, , ,
Citation(English)	IPSJ SIG Technical Report SLP, , ,
発行日 / Pub. date	2017, 7
権利情報 / Copyright	

口唇の深度画像を用いたディープオートエンコーダによるマルチモーダル音声認識

安井勇樹[†1]　岩野公司[†2]　井上中順[†1]　篠田浩一[†1]

概要：音声認識の雑音耐性を向上させる手段として，口唇の動画像情報を音声情報とともに利用するマルチモーダル音声認識の研究が数多く行われている．本研究では，Microsoft 社の Kinect で撮影された口唇の RGB 画像と深度画像を音声情報と共に利用するマルチモーダル音声認識手法を提案する．提案手法では，これら 3 つの特徴量を連結した高次元ベクトルを入出力とするディープオートエンコーダ（Trimodal Deep Autoencoder）を構築し，その中間層から得られる次元圧縮された特徴量を DNN-HMM に基づく音声認識の入力として利用する．20 名の話者による日本語文音声を用いて，白色雑音を 10dB で重畳した条件で認識実験を行った結果，音響情報のみを利用する場合と比較して，提案手法により単語正解精度が約 10%改善することが確認された．また，音声と RGB 画像の 2 つの情報を融合する場合よりも，深度情報を加えることで，正解精度が約 1%改善することも確認された．

キーワード：マルチモーダル音声認識，口唇，深度画像，ディープオートエンコーダ

1. はじめに

近年，音声認識技術を取り入れた実用システムの開発が数多く進んでいるが，周囲の雑音の影響によって認識性能が大きく劣化することが問題となっている．音声認識の耐雑音性を向上させるための一つの有効な手段として，音響情報だけでなく口唇の動画像情報を併用するマルチモーダル音声認識があげられ，様々な研究が行われている．

多くのマルチモーダル音声認識では，顔の正面から撮影された口唇画像を利用している[1-9]．しかし，発声時の口唇の形状が似ている音素対は，正面からの画像情報のみで正確に識別を行うことが難しい．そこで，奥行きを含めた 3 次元情報を用いる方法が考えられる．一つの方法として，顔の横方向から撮影された画像情報の利用が検討されている[10-12]．例えば，Kumar らは，2 台のカメラで顔の正面と横方向から同時に撮影を行うことで口唇を立体的に捉え，その情報を音響情報と併せて利用する音声認識手法を提案しており，その有効性を示している[12]．しかし，複数台のカメラが必要になることから，コストの高さや使用場所の制約の面で課題がある．

一方で近年，Microsoft 社の Kinect に代表される，被写体の 3 次元情報を容易に取得できるデバイスが安価に入手できるようになった．Kinect には通常のカメラの他に深度センサが搭載されており，RGB のカラー画像（以降，「色画像」と呼ぶ）と同時に，被写体の奥行きの情報を深度画像として獲得することができる．Kinect の登場により，口唇やその周辺領域の凹凸を 3 次元情報として安価・手軽に取得することができるようになり，Kinect を用いた自動読唇（リップリーディング）[13, 14] や，マルチモーダル音声認識[15-18]の研究が盛んに進められるようになった．本研究においても，口唇の 3 次元情報を Kinect を用いて抽出し，マルチモーダル音声認識に利用することを考える．

従来までのマルチモーダル音声認識では，画像情報と音響情報と融合する手法として，LDA などを利用した特徴量レベルでの融合やマルチストリーム HMM を利用したモデルレベルでの融合などが多く検討されてきた．それに対し，近年，様々な分野で高い有効性が確認されている「深層学習」を利用した融合手法が検討されるようになり，その有効性が確認されるようになった[19, 20]．Huang らは音響と口唇の色画像の 2 つの特徴量を入力とする DBN (Deep belief network) を構築し，その中間層から得られる融合特徴量を一般的な GMM-HMM 法に基づく認識の特徴量として利用する手法を提案しており，雑音環境下での有効性を確認している[19]．Ngiam らは音響と口唇の色画像から得られる 2 つの特徴量を入出力とするディープオートエンコーダ（Deep Autoencoder）を構築し，その中間層から得られる融合特徴量を利用した認識手法を提案している[20]．この手法による融合の有効性は示されているが，認識の枠組みとして SVM による線形識別を利用しているため，連続音声認識には対応していない．

以上の背景を踏まえ，本研究では，「音響情報」「口唇の色画像情報」「口唇の深度画像情報」の 3 者を，深層学習を利用して融合し利用する，マルチモーダル音声認識手法を提案する．具体的には，これら 3 つの特徴量を連結した高次元ベクトルを入出力とするディープオートエンコーダ（Trimodal Deep Autoencoder）を構築し，その中間層から得られる特徴量を DNN-HMM 法に基づく音声認識[21]の入力として利用する手法である．

以降では，まず 2 章で，提案する Trimodal Deep Autoencoder を用いたマルチモーダル音声認識手法の説明と，そのオートエンコーダの学習の方法について説明する．

†1 東京工業大学
Tokyo Institute of Technology
†2 東京都市大学
Tokyo City University

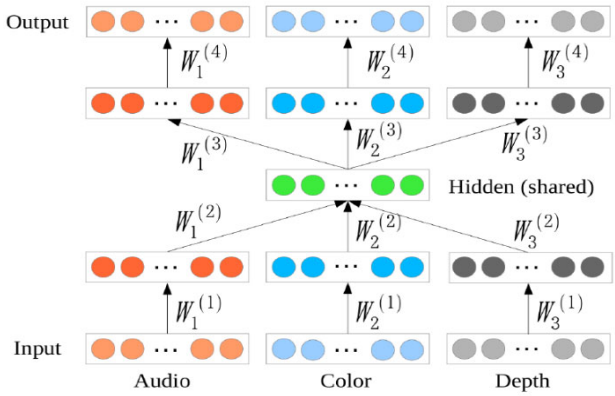

図 1　音響・口唇の色画像・口唇の深度画像の 3 情報を融合するためのディープオートエンコーダ
Figure 1　Trimodal Deep Autoencoder

3 章において，提案手法の評価に用いるデータと評価実験の条件，評価結果について述べ，最後に 4 章で本稿の結論を述べる．

2. Trimodal Deep Autoencoder

本研究では，Ngiam らが提案した音響と口唇の色画像の 2 つの特徴量をディープオートエンコーダで融合する方法[20]を，音響，口唇の色画像，口唇の深度画像の 3 つの特徴量の融合に拡張した手法を提案する．

提案するオートエンコーダ（Trimodal Deep Autoencoder）の様子を図 1 に示す．この図では 5 層のオートエンコーダの例を示している．入力は各モード（音響，色画像，深度画像）の特徴ベクトルであり，中間層でそれらを圧縮，出力層で再び各モードの特徴ベクトルが復元されるようなネットワークとなっている．共有されている中間層によって，各モードを融合した情報が表現されるため，この層の出力を新たな一つの特徴量として認識に利用する．ここで，$\mathbf{W}_i^{(l)}$ は i 番目のモードの l 層目のユニットの出力に対する（バイアスを含む）重みパラメータであり，$i = 1,2,3$，$l = 1,2,3,4$ となる．

$\mathbf{x}_i$ を音響（$i = 1$），口唇の色画像（$i = 2$），口唇の深度画像（$i = 3$）から得られる特徴ベクトルとするとき，それらにノイズを付与してオートエンコーダの入力とする．すなわち，それぞれのモードの入力の特徴ベクトル $\tilde{\mathbf{x}}_i$ は，q をノイズ付与関数として，$\tilde{\mathbf{x}}_i = q(\mathbf{x}_i)$ と表される．本研究では，関数 q を「確率 p で $\mathbf{x}_i$ の各成分をランダムに 0 に置き換える操作」としている．オートエンコーダの出力はノイズを付与する前の $\mathbf{x}_i$ とするため，元の入力が復元されるような Denoising Autoencoder [22] が構成されることになる．なお，事前学習，ファインチューニングの両方のプロセスともに入力の特徴ベクトルにはノイズ付与を行ったものを使用する．

図 2 に提案するディープオートエンコーダの事前学習のプロセスを示す．まず，それぞれのモードに対して個別にオートエンコーダを構成し，事前学習を行う（図 2(a)）．得られたオートエンコーダの上下の重みパラメータを分割し，目的とするディープオートエンコーダの出力側，入力側の層の初期パラメータに利用する．この例では，3 層で事前学習されたオートエンコーダの上下部分を分割し，$\mathbf{W}_i^{(1)}$ と $\mathbf{W}_i^{(4)}$ の初期パラメータとして利用している（図 2(b)）．その際に 3 モード分をまとめ，それらを共有する中間層を挿入してディープオートエンコーダを構成する（図 2(c)）．次に，中間部分（図 2(d)）を取り出し，3 モードの特徴ベクトルを連結した高次元ベクトルを利用して事前学習を行うことで，中間部分の重みパラメータ（$\mathbf{W}_i^{(2)}$，$\mathbf{W}_i^{(3)}$）の推定を進める．この事前学習におけるモード i の入力ベクトルは $\tilde{\mathbf{h}}_i$ と表され，以下の式(1), (2)で定義される．

$$\tilde{\mathbf{h}}_i = q(\mathbf{h}_i) \tag{1}$$
$$\mathbf{h}_i = \sigma(\mathbf{W}_i^{(1)}\mathbf{x}_i) \tag{2}$$

ここで，σ は活性化関数を表す．この 5 層のオートエンコーダの例では，ノイズが付与されていない入力に対する共有中間層の出力は，

$$\mathbf{h}' = \sigma\left(\sum_i \mathbf{W}_i^{(2)}\mathbf{h}_i\right) \tag{3}$$

と定義されるため，各モードに対する出力ベクトルは，

$$\mathbf{y}_i = \sigma(\mathbf{W}_i^{(3)}\mathbf{h}') \tag{4}$$

となる．このベクトルを中間部分の事前学習に用いる．

最後に全体に対してファインチューニングを行って，最終的な Trimodal Deep Autoencoder を得る．具体的には，以下の式 (5) の平均二乗誤差（MSE: Mean Squared Errors）で定義される損失関数を利用して重みパラメータを更新する．

$$\mathrm{MSE}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(y_{i,j} - x_{i,j})^2 \tag{5}$$

ここで，n_i はモード i の特徴ベクトルの次元数を表し，$y_{i,j}$ は $\mathbf{y}_i$ の j 番目の要素，$x_{i,j}$ は $\mathbf{x}_i$ の j 番目の要素である．

3. 評価実験

3.1 使用データ

提案手法の評価のため，20 名の話者による音響・画像データの収録を行った．発声内容は ATR 音素バランス文（503

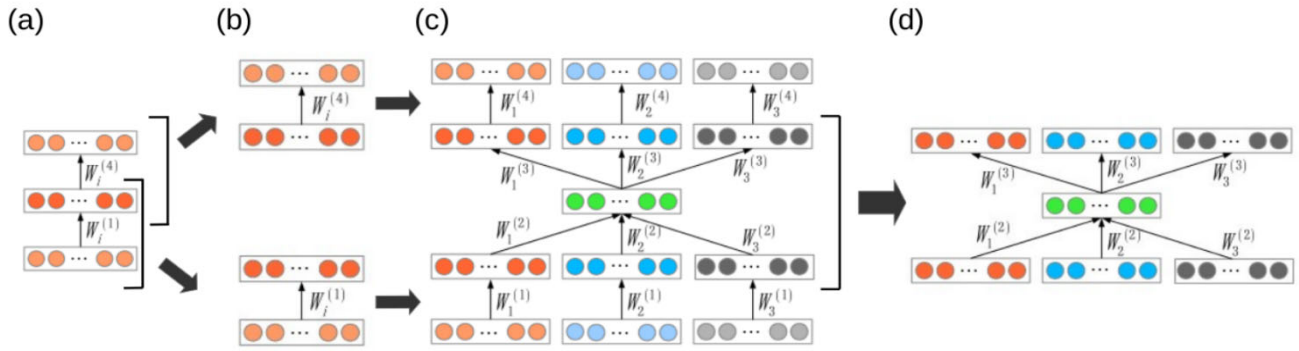

図 2　Trimodal Deep Autoencoder（5 層）の事前学習の流れ
Figure 2　Pre-training process of Trimodal Deep Autoencoder with five layers

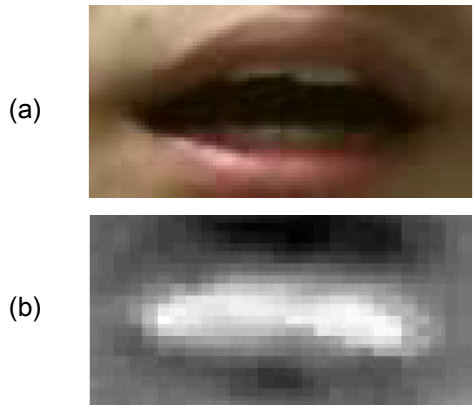

図 3　同じフレームにおける口唇の色画像(a)と深度画像(b)
Figure 3　A color mouth image (a) and a depth mouth image (b) captured at a same frame

文）[23]であり，各話者とも全 503 文を発声している．収録データ長は約 15 時間となった．

音響データは 44.1kHz, 16bit でサンプリングされており，特徴量抽出の前に 16kHz にダウンサンプリングしている．なお，収録は研究室で行ったため，キーボードのタイプ音や咳などの生活雑音を含んでいる．

画像データは Microsoft Kinect 2.0 で撮影を行った．フレームレートは 30Hz である．話者と Kinect の距離は固定していないが，深度情報を得るためには被写体がカメラから 0.5m 以上離れている必要があるため，それ以上の距離を保つように被験者には指示を行っている．各フレーム画像からの口唇領域の検出には，Microsoft 社が提供している開発キット Kinect for windows SDK 2.0 の機能を利用した．このキットによって顔の様々な特徴点位置を検出することができるので，上唇の上端，下唇の下端，唇の左端と右端の 4 点の座標を求め，それに基づいて口唇の矩形領域の切り出しを行っている．切り出し後の領域画像は 96×48 のサイズになるよう正規化している．なお，色画像と深度画像は座標系が異なっているため，同ツールで提供されている ICoordinateMapper クラスを用いて座標の補正を行っている．図 3 に同じフレームで観測された，正規化後の口唇の色画像(a)と深度画像(b)の例を示す．

3.2 実験条件

特徴量抽出を行うため，音響信号については 10msec ごとに 40 次元のメルフィルバンク係数を抽出する．グレースケール化された色画像と深度画像は，各フレームについて主成分分析に基づく次元圧縮を行い，それぞれを 32 次元のベクトルに変換する．音響のフレームレートは 100Hz, 画像のフレームレートは 30Hz と一致していないため，画像の特徴ベクトルを時間方向に 3 次元スプライン補間を施して 100Hz になるように補正する．ディープオートエンコーダへの入力／出力は，連続する 11 フレームの特徴ベクトルを連結したものとする．結果，音響の入力ベクトルは 440 次元，画像の入力ベクトルは 352 次元となる．

性能比較を行うため，以下の 4 条件のディープオートエンコーダ（DAE）に基づく特徴量化を行った場合の実験を行う．

- 音響のみについて，DAE に基づく特徴量抽出を行った場合（DAE_A）
- 音響と口唇の色画像について，Bimodal DAE で特徴量の融合を行った場合（BDAE_A_C）
- 音響と口唇の深度画像について，Bimodal DAE で特徴量の融合を行った場合（BDAE_A_D）
- 音響，口唇の色画像，口唇の深度画像について，Trimodal DAE で特徴量の融合を行った場合（提案手法：TDAE_A_C_D）

それぞれの場合で用いた，オートエンコーダの各層のユニット数を表 1 に示す．本研究では合計の層数を 7 とし，中央（4 番目）の層が共有化された中間層で，この層の出力を融合特徴量として認識に用いる．音声認識の音響モデルには triphone を単位とした DNN-HMM [21]を用いる．この DNN の隠れ層は 3 層で, それぞれの層のユニット数は 1,024

表 1　ディープオートエンコーダのユニット数
Table 1　The number of units in Deep Autoencoders

種類	モード	各層のユニット数
DAE_A	音響	440-200-120-80-120-200-440
BDAE_A_C	音響	440-200-120-80-120-200-440
	色画像	352-200-120-80-120-200-352
BDAE_A_D	音響	440-200-120-80-120-200-440
	深度画像	352-200-120-80-120-200-352
TDAE_A_C_D	音響	440-200-120-120-120-200-440
	色画像	352-200-120-120-120-200-352
	深度画像	352-200-120-120-120-200-352

とした．なお，DNN の入力は前後 5 フレームずつを連結した合計 11 フレーム分の特徴ベクトルである．言語モデルは毎日新聞記事データから学習された語彙数 63,465 の前向き 3-gram を用いる．認識性能は 4 分割の交差検証で評価し，学習データと評価データでは話者の重なりはない．雑音に対する頑健性を評価するため，評価データと学習データの音響信号に同じ SN 比で白色雑音を重畳した場合の実験も行った．評価は単語誤り率（WER: Word Error Rate）で行う．

ディープオートエンコーダの最適化手法には Adam[24] を用い，ノイズ付与の確率 p は 0.2，活性化関数 σ にはシグモイド関数を用いた．また，ミニバッチ学習におけるバッチサイズは 100 とした．なお，提案する Trimodal Deep Autoencoder の学習時間は，1 CPU（Intel Xeon X5670 2.93GHz），1 GPU（NVIDIA Tesla K20X 1.31 Tflops）のマシンで約 48 時間であった．

音響情報のみを利用した場合のベースラインの認識性能を検証するため，ディープオートエンコーダを使用せずに，各フレームから 12 次元の MFCC と対数パワーを抽出して 11 フレーム分を連結し，DNN-HMM の枠組みで認識を行う手法（MFCC_A）による性能評価も行った．

3.3 評価結果

表 2 に実験結果を示す．白色雑音は 10, 20dB で重畳を行っている．クリーン条件における MFCC_A と DAE_A の結果を比較すると，オートエンコーダによる特徴抽出の導入により若干の性能劣化が生じていることがわかる．一方で，雑音が重畳された条件では，オートエンコーダによる雑音除去の効果が得られ，その差が小さくなっていることがわかる．

口唇の色画像，深度画像のそれぞれを音響情報と融合した場合の結果（BDAE_A_C, BDAE_A_D）の結果を見ると，DAE_A に比べ，クリーンや SN 比 20dB の条件では大きな性能改善は得られていないが，SN 比が 10dB の条件で大きな性能改善がみられ，単語正解精度がそれぞれ 9.8%, 11.1% 向上していることがわかる．

表 2　認識実験の結果（単語誤り率 %）
Table 2　Experimental results (Word Error Rates %)

実験条件	SN 比		
	clean	20dB	10dB
MFCC_A	20.3	23.7	34.6
DAE_A	21.8	24.6	35.0
BDAE_A_C	21.9	23.4	25.2
BDAE_A_D	21.1	22.3	23.9
TDAE_A_C_D	20.9	22.4	24.2

次に，提案手法の結果（TDAE_A_C_D）を見ると，雑音環境下では音響情報のみの結果（MFCC_A, DAE_A）よりも有意に単語誤り率の削減が得られ，SN 比 10dB の条件では約 10%の性能向上を示している．また，色画像情報を融合した場合の結果（BDAE_A_C）と比べると，深度情報の融合により 1%程度の性能向上が得られていることがわかる．ただし，提案手法の結果（TDAE_A_C_D）が深度情報のみを融合したときの結果（BDAE_A_D）とほとんど変わらない結果であることから，3 者の融合による効果がまだ十分に得られていないことがわかる．

3.4 考察

今回の実験結果を見ると，口唇の色画像よりも深度画像の情報の方が耐雑音性の向上に有効に作用していることがわかる．その要因としては，深度情報の方がより正確に口腔内の状況(舌や歯の状態など)を反映することができる，深度情報の方が話者の違いに対して普遍性が高く，頑健である，などが考えられる．今回は PCA による特徴量化を行っているが，より正確な特徴量抽出を行うため，SIFT や HOG といった局所特徴量の利用や，DCT による周波数成分の利用などを検討する必要がある．

SN 比 10dB における DAE_A と BDAE_A_C の認識結果（出力テキスト）の比較を行ったところ，色画像情報の融合によって，口を閉じて発声される母音や子音(/m/ など)に対する認識誤りが削減されていることが確認された．また，BDAE_A_C と TDAE_A_C_D の結果を比較すると，深度情報の導入により，/k/, /t/ といった発声時の口唇の外見だけでは判断が難しい音素対の識別性能が向上していることがわかった．図 4, 5 は，それぞれ同じ母音に続く /k/, /t/ を発声したときの口唇の色画像(a)と深度画像(b)を示している．色画像では両者にはっきりとした違いが見られないが，口腔内の舌の位置が異なっているため，深度画像では両者の特徴に違いが表れていることがわかる．

4. おわりに

本研究では，音声情報と口唇の色画像情報，深度画像情報の 3 者をディープオートエンコーダ（Trimodal Deep Autoencoder）によって融合するマルチモーダル音声認識手

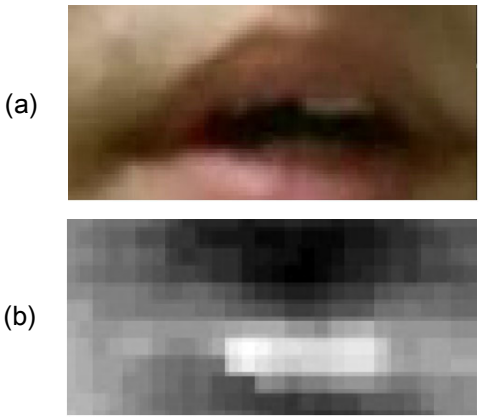

図 4　/k/ を発声したときの口唇の色画像(a)と深度画像(b)

Figure 4　A color mouth image (a) and a depth mouth image (b) when /k/ is pronounced

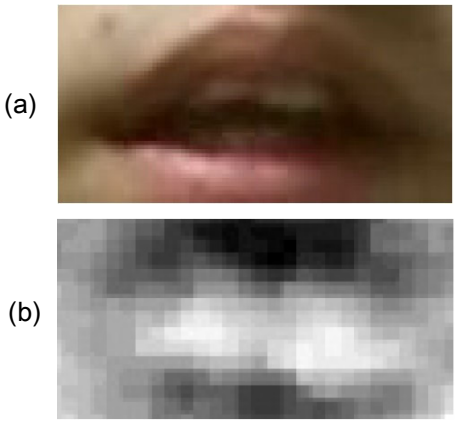

図 5　/t/ を発声したときの口唇の色画像(a)と深度画像(b)

Figure 5　A color mouth image (a) and a depth mouth image (b) when /t/ is pronounced

法の提案を行い，その有効性の検証を行った．色画像，深度画像の撮影には Microsoft 社の Kinect 2.0 を用いている．20 名の話者による日本語文音声を用いた実験の結果，白色雑音を 10dB で重畳した条件で，音響情報のみを利用する場合と比較して，提案手法により単語正解精度が約 10%改善することが確認された．また，音声と色画像の 2 つの情報を融合する場合よりも，深度情報を加えることで，正解精度が約 1%改善することも確認できた．

今回の実験では，音響信号に対する雑音成分のみを考慮しているが，今後は画像信号に対する雑音成分（例えば，カメラの角度の変化など）についても考慮した評価が必要となる．また，現状では音響信号の雑音として白色雑音のみの評価に留まっているため，他の様々な種類の雑音に対する頑健性の評価を行う必要もある．認識性能の改善に向けては，層数やユニット数など，各種の実験パラメータの調整（最適化）を進めることや，教師ありの深層学習に基づく情報融合手法の検討などを行う必要がある．

謝辞　この研究は科研費 16H02845 の支援を受けた．

参考文献

[1] C. Bregler and Y. Konig, ““Eigenlips” for robust speech recognition,” Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’94), vol. 2, pp. 669–672, Adelaide, SA, Australia, 1994.

[2] M. J. Tomlinson, M. J. Russell, and N.M. Brooke, “Integrating audio and visual information to provide highly robust speech recognition,” Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’96), vol. 2, pp. 821–824, Atlanta, GA, USA, 1996.

[3] G. Potamianos, E. Cosatto, H. P. Graf, and D. B. Roe, “Speaker independent audio-visual database for bimodal ASR,” Proc. ESCA Workshop on Audio-Visual Speech Processing (AVSP ’97), pp. 65–68, Rhodes, Greece, 1997.

[4] C. Neti, G. Potamianos, J. Luettin, et al., “Audio-visual speech recognition,” Final Workshop 2000 Report, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, USA, 2000.

[5] S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” IEEE Transactions on Multimedia, vol. 2, no. 3, pp. 141–151, 2000.

[6] Y. Zhang, S. Levinson, and T. S. Huang, “Speaker independent audio-visual speech recognition,” Proc. IEEE International Conference on Multi-Media and Expo (ICME ’00), pp. 1073–1076, New York, NY, USA, 2000.

[7] S. M. Chu and T. S. Huang, “Bimodal speech recognition using coupled hidden Markov models,” in Proc. the 6th International Conference on Spoken Language Processing (ICSLP ’00), vol. 2, pp. 747–750, Beijing, China, 2000.

[8] C. Miyajima, K. Tokuda, and T. Kitamura, “Audio-visual speech recognition using MCE-based HMMs and model-dependent stream weights,” Proc. the 6th International Conference on Spoken Language Processing (ICSLP ’00), vol. 2, pp. 1023–1026, Beijing, China, 2000.

[9] K. Iwano, S. Tamura, and S. Furui, “Bimodal speech recognition using lip movement measured by optical-flow analysis,” Proc. International Workshop on Hands-Free Speech Communication (HSC ’01), pp. 187–190, Kyoto, Japan, 2001.

[10] T. Yoshinaga, S. Tamura, K. Iwano, and S. Furui, “Audio-visual speech recognition using lip movement extracted from side-face images,” Proc. Auditory Visual Speech Processing (AVSP), pp. 117–120, St. Jorioz, France, 2003.

[11] P. Lucey and G. Potamianos, “Lipreading using profile versus frontal views,” Proc. IEEE Multimedia Signal Processing Workshop, pp. 24–28, 2006.

[12] K. Kumar, T. Chen and R. M. Stern, “Profile view lip reading,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), vol.4, pp.429–432, Honolulu, HI, USA, 2007.

[13] A. Yargic and M. Dogan, “A lip reading application on MS Kinect camera,” Proc. IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA), Albena, Bulgaria, 2013.

[14] A. Rekik, A. Ben-Hamadou, and W. Mahdi, “An adaptive approach for lip-reading using image and depth data,” Multimedia Tools and Applications, vol. 75, iss. 14, pp. 8609–8636, 2016.

[15] G. Galatas, G. Potamianos, and F. Makedon, “Audio-visual speech recognition incorporating facial depth information captured by the Kinect,” Proc. IEEE European Signal Processing Conference (EUSIPCO), pp.2714–2717, Bucharest, Romania, 2012.
[16] 押尾翔平, 岩野公司, 篠田浩一, “口唇の深度画像を用いたマルチモーダル音声認識，” 情報処理学会研究報告，vol.2014-SLP-102, no. 2, pp. 1-6, 2014.
[17] K. Palecek, “Comparison of Depth-Based Features for Lipreading,” Proc. IEEE International Conference on Telecommunications and Signal Processing (TSP), Prague, Czech Republic, 2015.
[18] J. Wang, J. Zhang, K. Honda, J. Wei, and J. Dang, “Audio-visual speech recognition integrating 3D lip information obtained from the Kinect,” Multimedia Systems, vol.22, pp. 315–323, 2016.
[19] J. Huang and B. Kingsbury, “Audio-visual deep learning for noise robust speech recognition,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), pp.7596–7599, Vancouver, BC, Canada, 2013.
[20] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” Proc the 28th International Conference on Machine Learning (ICML-11), pp. 689-696, Bellevue, WA, USA, 2011.
[21] F. Seide, G. Li, and D. Yu, “Conversational speech transcription using context-dependent deep neural networks,” Proc. 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), pp. 437–440, Florence, Italy, 2011.
[22] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” Journal of Machine Learning Research, vol. 11, pp. 3371–3408, 2010.
[23] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol. 9, pp. 357–363, 1990.
[24] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Proc. International Conference for Learning Representations (ICLR), San Diego, CA, USA, 2015.